

# NIKITA MARKOV

Email : [nikitamarkov.work@gmail.com](mailto:nikitamarkov.work@gmail.com)  
Blog : [desire32.github.io/blog/](https://desire32.github.io/blog/)  
LinkedIn : [Nikita Markov](https://www.linkedin.com/in/nikitamarkov/)

Phone : +357 97 533793  
Github : <https://github.com/Desire32>

## ABOUT ME

Machine Learning Engineer with hands-on experience in ASR, NLP, and LLM fine-tuning, combining research publications, real-world projects, and strong engineering skills.

## TECHNICAL SKILLS

**Programming Languages:** Python (3.10+), C++, Go (<1.23)  
**Machine Learning / Deep Learning:** PyTorch, Transformers, scikit-learn, NumPy, Pandas  
**NLP / LLMs:** Hugging Face, LangChain, LoRA, RAG, ASR  
**Data & Infrastructure:** PostgreSQL, Redis, Kafka, Docker, Git, AWS (EC2, RDS)  
**MLOps & Monitoring:** MLflow, Weights & Biases  
**Systems & Tools:** Linux, Unix, SSH, ONNX, TensorRT  
**Editors / Workflow:** LazyVim, Zed

## EDUCATION

**University of Central Lancashire** Cyprus  
BSc in Computer Engineering / Computing Grade: First Class | 09/22 – 09/26  
*Thesis:* Performance Evaluation of UE-VBS as Computational and Storage Hub CSHs in 6G Networks with RL integration

## WORK EXPERIENCE

**RIF Internship — Abasis AI** Jul 2025 – Aug 2025  
Cypriot ASR dialect model, [News post](#)  
– Developed an ASR model from scratch in 6 weeks, based on the Wav2Vec2 architecture.  
– Integrated KenLM language model as an intermediate module, improving WER (word error rate) by 7%.  
– Prepared and curated training data using pandas and yt-dlp  
– Trained and fine-tuned models on the brev nvidia.

**InSPIRE Research Center Research Assistant** Oct 2024 - Present  
**Frugal AI Techniques for LLM Deployment on NVIDIA Jetson Orin Nano — UNPUBLISHED**  
– Reduced LLM memory footprint from 1.1B to 470M parameters using Frugal AI techniques, enabling deployment on edge hardware without degradation in response quality.  
– Built a TensorRT-accelerated inference stack on NVIDIA Jetson Orin Nano leveraging ONNX export, trtexec optimization, and containerized execution via jetson-containers (NanoLLM).  
– Developed a modular LLM runtime supporting dynamic model/embedding selection, multiple quantization schemes (INT3–INT8), and RAG-based retrieval with detailed runtime telemetry.

**Enhancing Digital Heritage Experiences: Evaluating Fine-Tuned LLM Integration — Publication**  
– Developed a modular fine-tuning pipeline supporting multiple architectures (TinyLlama, Mistral-7B, Llama-8B, Phi-2) with custom qLoRA configurations for efficient training  
– Implemented a sophisticated dual-pipeline architecture combining FAISS-based semantic search and Word2Vec embeddings, featuring dynamic context retrieval and optimized text chunking for enhanced knowledge access  
– Integrated MLflow for comprehensive experiment tracking, model versioning, and performance metrics visualization across different model architectures and training configurations

**Developing a Cyber-Physical-Social Metaverse System for Cultural Experiences — Publication**  
– Accepted into a peer-reviewed research paper with a competitive 22% acceptance rate.  
– Architected a sophisticated chatbot system using LangChain, integrating local LLM inference via Ollama with Mistral-7B model and persistent memory storage using Upstash Redis.  
– Implemented a scalable Flask-based REST API with streaming response support and session-based chat history management, enabling seamless conversation persistence across multiple user sessions.  
– Deployed the system on AWS EC2, demonstrating production-ready architecture with proper environment management and security considerations.

## LANGUAGES

English (C1), Russian (Native)